

# Genomics & Proteomics™

January/February 2005 ♦ Vol. 5, No. 1

Reed Business  
Information

*Trends and Technologies Fueling Omics Research*

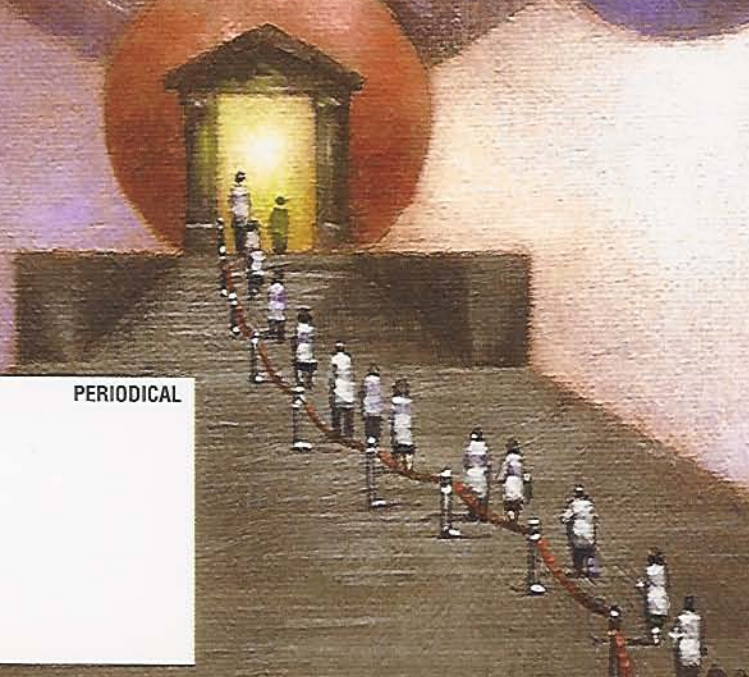
## Molecular Libraries' Grand Opening

*NIH brings academic researchers tools  
once only available to big pharma*

- **PCR Questions Linger**
- **High Variability in Gene Expression**
- **Microarrays Leave Lab**

[www.genpromag.com](http://www.genpromag.com)

PERIODICAL





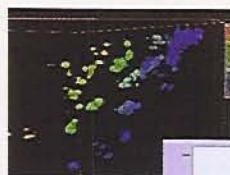
# Contents

## Features

- 14** COVER: Big Plans for Small Molecules
- 20** PCR Leaves its Teen Years, and Lingering Questions, Behind
- 26** NMR Is Ready, so Bring on the Macromolecules
- 30** Tackling High Variability in Gene Expression Studies
- 33** Microarrays Get Ready to Step Outside the Laboratory

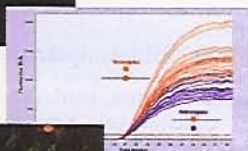
## Departments

- 5** Staff
- 7** Editorial
- 9** Editorial Index
- 9** Advertiser Index
- 11** Transcription
- 13** Genomics & Proteomics World Report



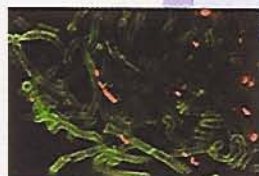
**26**

Nuclear Magnetic Resonance



**20**

PCR Advances



**33**

Array Technology

"The [molecules] that get thrown out in the [pharma industry] trash could be the ideal starting place for someone working in a related but different field."

— Douglas Livingston, PhD, senior vice president of chemistry, Discovery Partners International

**14**

**COVER STORY** As part of its Roadmap for Medical Research, NIH is embarking on initiatives that will equip academic researchers with the same chemical tools and molecular libraries available to most pharmaceutical companies. The effort, called the Molecular Libraries Initiative, is likely to have a profound influence on the understanding of biological pathways. (Illustration by Paul Anderson)

## GET INTO SOMETHING NEW:

Every Time You See This Symbol:

**InfoLINK**

Detailed product specifications, applications data, and purchasing information are available...

**IMMEDIATELY.** See page 5 for details.



# Big Plans for

■ *As part of its Roadmap for Medical Research, NIH is embarking on a sweeping series of initiatives to equip academic researchers with the same toolset available to most pharmaceutical companies. The effort, called the Molecular Libraries Initiative, is likely to have a profound influence on our understanding of biological pathways.*

■ **By Bill Schu, Senior Editor**

What if 3M had decided that their new glue just didn't work, and simply tossed aside the technology that ultimately led to the Post-It note? How would we jot down people's phone numbers and remember to call them back?

What if George de Mestrel had simply ripped the burrs off his clothing and thought nothing more of them, rather than studying how they attached to his coat and ultimately inventing Velcro? What if the Kellogg brothers had been a little more attentive and hadn't left their cooked wheat untended for several days, producing cornflakes?

The "success stories" above are told and retold, classic examples of serendipity. They are notable and memorable partly because of the human enjoyment of the unlikely, partly because they seem to reflect that pure chance could happen to any one of us. What if we had only done this, or hadn't done that?

The what-ifs that haunt life sciences research are in a category all their own: What if that small molecule that didn't show much promise could have led to a cure for breast cancer? What if a small molecule that interacts with a gene and shows some toxicity could interact with some proteins in a biologically beneficial manner?

Christopher Austin, MD, senior advisor for translational research for the National Human

Genome Research Institute, part of the National Institutes of Health (NIH), has perhaps been more haunted by these possibilities than most. In his previous position as a senior researcher at Merck & Co., he was familiar with the kinds of information that large pharmaceutical companies had access to, but that academic researchers did not. Of particular interest to Austin was information surrounding how small molecules interacted with proteins, DNA, and RNA. NIH is banking on the fact that academic researchers, if given access to heretofore unavailable libraries of organic chemical compounds or small molecules, can use those molecules to better understand the biological pathways associated with human health.

"What was frustrating at Merck was that unless the target and/or the small molecule against that target had rather short-term commercial potential, no matter how interesting the biology was, it had to be abandoned," Austin says. "Roughly 95% of science is not going to lead to a short-term commercial result. I ended up feeling like a kid in a candy shop, but I couldn't touch the candy. I was surrounded by these wonderful tools and methods and I couldn't access them."

One person's target is another's anti-target, says Douglas Livingston, PhD, senior vice president of chemistry at Discovery Partners International (DPI) Inc., San Diego, which was awarded a multiyear contract to set up and maintain the small-molecule repository to manage and provide up to



# Small Molecules

1 million chemical compounds for NIH and academic screening centers that will be selected in 2005. "The stuff that gets tossed out in the trash could be the ideal starting place for someone working in a related but different field."

Austin and his colleagues at NIH decided to solve their question: What if academics studying biological pathways had access to information about small-molecule interactions? "Pharma had already figured out how to do this on an operational level," says Austin. "What we needed to do was convince the public sector to invest in this infrastructure and the human capital to rigorously apply these methods to problems of basic biology. I firmly believe that better understanding of basic biology will underpin better therapeutics. To the degree that we can define better biology, gene function, network function, and pathways, we'll improve the predictive effects of small molecules in living systems over the long term."

Now, the NIH Chemical Genomics Center (NCGC) is poised to answer those questions. A coordinated effort will include at least four distinct groups within NIH, a handful of academic screening centers, and the creation of PubChem, a new database that is to house hundreds of thousands of molecular structures. Many officials within NIH believe it is the most ambitious undertaking the group has embarked on since the completion of the map of the human genome.

## Showing some initiative

Traffic on Wisconsin Avenue in Bethesda, Maryland, the site of NIH headquarters, hasn't gotten any lighter since the recent doubling of the NIH budget. No existing roadmap, not even NIH's own Roadmap for Medical Research, has had any luck solving that problem. Of course, easing snarled traffic isn't the goal of the NIH

Roadmap; the goal is nothing less than the complete transformation of the nation's medical research capabilities. As part of these efforts, NIH plans to address critical roadblocks and knowledge gaps that currently constrain rapid progress in biomedical research.

The NCGC, formed in 2003, is the first component of a nationwide network that will produce innovative chemical tools for use in biological research and drug development.

Illustration by Paul Anderson



The next components are the Molecular Libraries Initiative (MLI) and the Molecular Libraries Screening Centers Network (MLSCN). As many as six extramural chemical genomic pilot centers will be funded at academic institutions across the country to screen assays submitted by the research community on a large number of compounds maintained in a central compound repository. The small-molecule library will initially consist of some 100,000 chemically diverse small molecules with both known and unknown activities.

"The intent of doing the genome project in the first place was to derive biological insights and therapeutic potential out of the genome," says Austin. "It was clear to us that [the MLSCN] would be a good tool for doing that. Most of NIH is categorical, set up to work on particular disease areas, particularly areas of orphan diseases that don't have enough of an economic prevalence to get private industry interested. Those disorders might benefit both from the provision of small-molecule tools to study the physiology of genes that are affected in the disorders and also perhaps serve as a starting point to development of an actual therapeutic."

"The first year has been occupied by acquiring and curating the first 100,000 molecules," says Livingston. "Starting in the next few years, as we acquire the second, third, and fourth sets of compounds, each set is going to inform the next. We're going to have access to all of the screening data, which will be in the public domain. That's going to tell us where the richest biology is going to be and inform us as to which compounds should go into the collection."

### Mother knows best

What might screening of hundreds of thousands of small molecules accomplish, beyond what pharmaceutical companies that have screened small

molecules intensely for many years have been able to do?

For one thing, says Austin, the MLSCN will explore the vast majority of the human genome for which no small-molecule chemical probes have been identified. Of the hundreds of thousands of proteins thought to be encoded by the 25,000 genes in the human genome, less than 500 currently have a chemical compound with which they interact. If researchers can discover how small molecules will interact with other proteins, DNA, and RNA, they should learn more about how disturbances in intricate biological pathways lead to disease, and what can be

done to better treat, or even prevent, certain diseases.

"As it became evident to the genome project that the way Mother Nature generates human complexity is not by increasing the number of genes but by increasing the number of proteins, then it became in my mind more and more imperative to start working on the level that Mother Nature works," says Austin. "Mother Nature works at the level of proteins, and if you understand what the genome does functionally, you have to work at the level at which genes lead to traits. DNA or gene locus rarely leads to a single trait. A gene will lead to multiple RNAs, which will lead

## Blueprint for Small-Molecule Analysis

Beyond NIH's efforts to form the Molecular Libraries Initiative, others are developing parallel efforts to enhance the utility of small-molecule screening. One such group is the Blueprint Initiative, a research program of the Samuel Lunenfeld Research Institute (SLRI) at Mount Sinai Hospital, Toronto, Canada. Under the leadership of Christopher Hogue, PhD, principal investigator, the Blueprint Initiative is developing, hosting, and maintaining public biological databases and bioinformatics software tools—and providing them to academic researchers for free. Blueprint's Small-Molecule Interaction Database (SMID), links small molecules to their protein partners and families. *Genomics & Proteomics* talked to Hogue about what he hopes SMID will accomplish.

### G&P: Can you briefly describe SMID?

**Hogue:** SMID is a mechanism to assign potential small-molecule binding sites to sequences of unknown function. It was conceived as a drug discovery tool to analyze sequences, but it is being delivered as an open-source, freely available tool. It works by analyzing all the small-molecule crystal structures in the Protein Data Bank and their contacts to proteins.

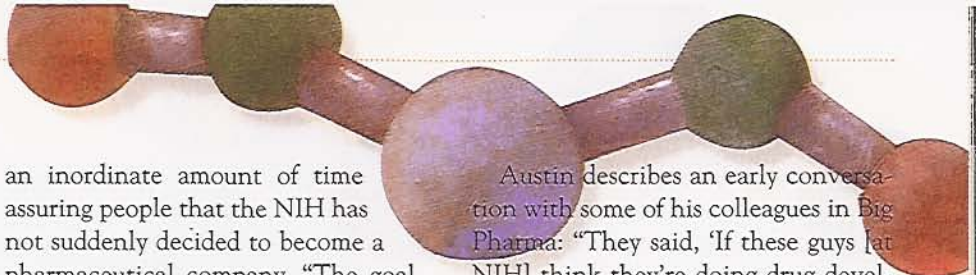
### G&P: What does the analysis entail?

**Hogue:** We first assign domains to the sequence, and then we dip into this relational database that we've constructed. We pull out the specific binding sites to each sequence, and we look at the columns in the sequence alignment that have the exact contacts to the small molecule. Rather than a domain that may be 50 to 70 amino acids [in length], we're only looking at maybe six to 12 amino acids in the context of that domain. We look at whether or not those are the same as the residues that are in a crystal structure. If we can assert that, then we can say that it has the same fold, the same domain, and then we have something.

### G&P: How is SMID different from what NIH is doing with the MLI and PubChem?

**Hogue:** PubChem is a collection of small molecules that [NIH] is trying to put into a GenBank-like format. They're also trying to set up essentially data warehouses of the results of high-throughput screens of various types of chemical and genetic screens.





to multiple proteins, which will lead to a trait.

"If you take the general principles of pharmacology—that is, that small molecules have an ability to interact potently and sometimes specifically with molecular targets and induce a specific physiological response on the level that Mother Nature does biology—you can take those principles and apply them to problems of basic research instead of problems of therapeutics. Then you understand what we're up to."

#### **Not embarking on drug discovery**

Since the MLI and MLSCN were introduced, NIH officials have spent

an inordinate amount of time assuring people that the NIH has not suddenly decided to become a pharmaceutical company. "The goal from the beginning was never to make drugs," says Austin. "People have preconceptions about how you can use small molecules, based on how they've always been used before."

"The MLSCN has a strong focus on those areas where drug discovery drops off," says Livingston. "One thing that is missed sometimes is that we're intending to put toxins into the library—intentionally—which really distinguishes [this effort] from classical drug discovery."

Austin describes an early conversation with some of his colleagues in Big Pharma: "They said, 'If these guys [at NIH] think they're doing drug development on a budget of 50 million or 100 million dollars a year, they're [dreaming].'"

The confusion arose initially because small molecules have been used mostly in the private sector for doing drug development. "Once you throw the word 'screening' in for small molecules," says James Inglese, PhD, director of biomolecular screening and profiling at NCGC, "then the dynamic changes and it's associated with drug development. People equate screening of small molecules with drug discovery."

But, according to Austin, that's an independent question from what small molecules could be used for. What if they could be used as a screening tool in academic setting, as a way of better understanding certain signaling pathways? Only a handful of institutions in the public sector, such as Harvard, have screened small molecules. Otherwise, it has been limited to the private sector for the purpose of identifying leads for drug development. "Screening is on the way to drug development," Austin says. "But the science simply is much, much richer than that. The community has finally begun to appreciate the potential for this approach to biology."

"I am not aware of anybody discovering a drug directly out of synthetic, small-molecule libraries such as this," says DPI's Livingston. "If anything, we may hand off a better validated target or a better understanding of where to go for carrying it forward. The compound itself is overrated in terms of its importance in the overall process. That's not to say that one can't sometimes get lucky. With the prodigious amount of screening that will go on with this collection and the amount of scrutiny towards it, it has a higher probability than normal. It might not be as high as 1 in a billion, but I don't

I don't know if it really provides any further sequence analysis tools. I know that they've done some work in creating a structure database, but I don't think it looks at the interactions at a sequence level between specific residues and small molecules. While NIH may not have specific information about what proteins a small molecule binds to, they may have more phenotypic information on a small molecule, in combination with a knockout, for example. I see people skipping between these resources quite freely.

#### **G&P: Will there be opportunities for collaboration between Blueprint and NIH?**

**Hogue:** Steven Bryant is organizing PubChem, and I was a postdoc in his lab in the 1990s. We have a tight collaboration with NCBI on the small-molecule front. I wouldn't say we're doing anything better or different; I would just say that we've taken on different hard tasks to chew on and we're working well together on those.

#### **G&P: What benefits can you foresee arising out of the use of SMID?**

**Hogue:** Sometime this year, we will be able to provide scientists with an interface where they can take a list of genes, say 60 or 100 genes that they find up-regulated or down-regulated in any kind of a disease process by a microarray. They will be able to take that set of results and map it to a predicted protein interaction map and superimpose all the small-molecule binding sites of that protein interaction pathway. That will take the data out of microarrays and present it in more of a mechanistic format.

It will give scientists the ability to devise simple experiments by perhaps putting inhibitors for those small molecule binding sites into an organism and getting very quickly to some experimental designs to help sort out the mechanisms they see through microarray data. Tying microarray data together with protein interaction and small-molecule binding sites can really have a powerful effect on discovering ways of modulating pathways with multiple small molecules. The ability for scientists to map the results into mechanisms will take a lot of information that has been piling up on the shelves about lists of genes that are implicated in one biological process or another.



think anyone is holding their breath."

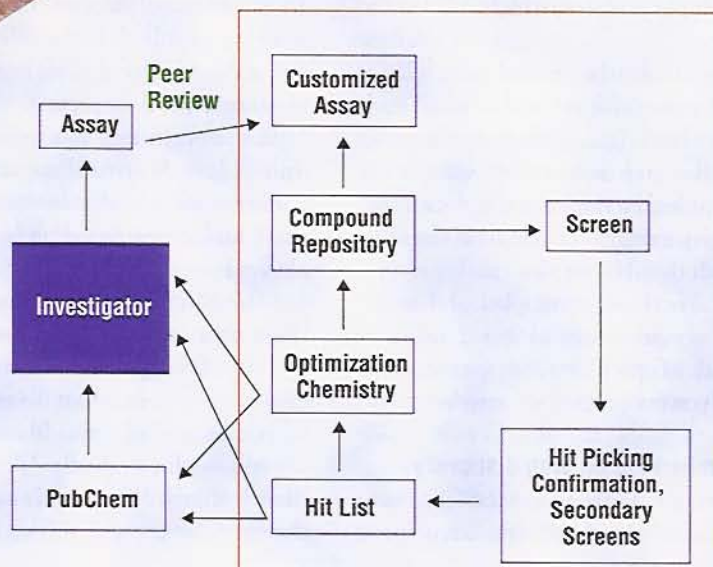
"If [the molecules] look promising for hitting novel parts of biological space, they will be expanded into larger libraries by the repository budget, and then put into the repository and screened on a larger basis by the screening centers," says Austin. "The assay development will lead to screens that are done in the screening centers. All of those results will be published in the literature and on PubChem."

### Birth of a database

In October 2003, Christopher Austin and his team came to Steve Bryant, PhD, senior investigator with NIH's National Center for Biotechnology Information (NCBI), with what seemed like an impossible task: make a repository and information retrieval system where the results of the MLSCN can be stored and categorized for use by academic researchers. That wasn't even the difficult part. "By when do you need it?" Bryant asked his bosses. Of course, he says, the answer was: "Yesterday."

"I'm a protein crystallographer by background," Bryant says. "When I started thinking about the task, on the one hand, it was very new, involving chemical structures and assay data. But on the other hand, it wasn't hard to translate the technical approach we've used with other databases. We weren't literally recycling other pieces, but we were doing things that were similar."

Bryant and his colleagues cobbled together the database, called PubChem, out of new and used parts—a piece borrowed here from the National Library of Medicine, a piece borrowed there from the national literature database PubMed. "The technical part, the informatics, is knowing how to create valence bond



As part of the MLSCN, molecular data will flow freely into and out of PubChem. (Source: Christopher Austin, MD, National Institutes of Health)

models from these very heterogeneous inputs that follow a set of standard rules so that we can automatically group the compounds by similarity and identity," says Bryant. "That's one of the most important retrieval properties we knew we would need."

But the database also needed to hold assay data, a task for which there was no model from which Bryant and his team could borrow. "We had to come up with a one-size-fits-all way to store and describe assay data," says Bryant. "We also had to come up with a retrieval system where people could look at the readouts and descriptions and retrieve compounds and compare them for structural similarity."

Although NCBI officials knew the screening centers for the MSLCN would not be identified until the spring of 2005, they set a goal of October 2004 for completion of PubChem, so that researchers at the screening centers could see how the database would look when they started to design their and implement their assays.

Bryant envisions a host of benefits that will arise from having a robust Pub-

Chem database freely available to the academic community. "It would be useful to have small-molecule structures in the NCBI collections so you can look up lists of related structures and the papers that go with them, just like you could look up genes and the papers that tell you what they do," he says. "The power of the molecular database, whether it's protein sequences or protein structures, is that computers are great at comparing these things. And people are terrible at giving [molecules] names that make sense. I read recently that the first international conference on standardizing chemical nomenclature was in 1892. They still don't have it right."

While bioinformatics was the hot field in the late 1990s during the genome project, Austin says that cheminformatics is the next big thing. "The reason that bioinformatics was so important was that the public efforts were generating enormous amounts of raw data and putting it into the public domain, which could then be mined for patterns across species or within species, looking for related structures," says Austin. "In this case, DNA sequence.



Apply that same principle now to small molecules interacting with targets. For the first time, there is going to be an enormous data set that will allow people to associate small-molecule structures with biological structures and protein structures, or physiology. And it will get richer as time goes on."

### Scavenging for treasure

It is no coincidence that Austin and Inglese have a background in Big Pharma. Hand in hand with their experience at Merck came an intimate knowledge of both the powerful tools at their disposal and a nagging lament about their inability to use some of those tools to their fullest public health potential. But don't blame Big Pharma for what it doesn't do. Austin says that researchers at pharmaceutical companies realize and appreciate that the science they can't necessarily use is good.

"The level of science that goes on at the pharma companies I've seen is exceedingly high—equal to what goes on in academia," says Austin. "It's not that people don't realize that the problems are out there. The problem is that the market is a very tough taskmaster. The market requires that you abandon all but the most immediately promising science."

Now, with a completely different goal, Austin and Inglese will be picking up where drug discovery often leaves off. "Will we be making drugs ourselves? Of course not," Austin says. "But does that mean that this activity will have no effect on therapeutic development? I hope it does have an effect, in the same way as NIH research has always done."

"I think MLSCN will be important in helping the pharma industry get new

drugs, look at new mechanisms," says Livingston. "It's been said that blockbusters are going to come from new biology, not new chemistry. I think that's absolutely true. A big part of this, of course, is systems biology. Using these small molecules is often an ideal tool to use as molecular probes. They can be very powerful in biology as well."

In the short term, says Austin, the MLI will allow researchers to have their assay taken in by the network, and be delivered a potent and soluble inhibitor or activator of that process. "For people

"[Researchers] are taking kind of a wait-and-see approach, because they don't know what kind of information is going to be there. No one knows yet. The screening centers are just being reviewed, and there's a peer review process anticipated for how assays get into the centers and for how compounds get accepted to the central repository."

For his part, Austin is ready to grab some of that previously out-of-reach candy. "This is the most exciting thing I've ever been involved with," he says. "In terms of being catalytic to further

### GLIMPSE THE FUTURE

*"The one thing I would love to see in five years is not so much where we're at, but the rate. This is one of those unique cases where the 5,000th piece of information is more valuable than the 5th. We saw that, for example, in the assembly of the genome, how it suddenly accelerated and snapped itself together, which was actually predicted by mathematicians. I think we're going to start to see the understanding go up exponentially."*

— Douglas Livingston, PhD, senior vice president of chemistry at Discovery Partners International

who never submit an assay via PubChem who are interested in using these kinds of tools to further their research, this is a complementary tool that works at the protein level," says Austin.

### Forming the "functionation" toolbox

Austin ultimately hopes the MLSCN becomes an integral part of life science research. "Think about the fact that people can get into GenBank and have not only the gene sequence, the protein sequence, the homologies with other organisms, the polymorphisms," he says. "But you'll also have a link to where to get a knockout mouse, an antibody, and a small molecule. That's a fundamentally different toolbox—a 'functionation' toolbox."

NIH officials are openly optimistic that the MLI will become a crucial part of the research toolbox. But NCBI's Bryant acknowledges that, for the most part, the academic research community is holding off until they see some results.

development, we've actually designed the initiative to have various parts that interlock, in order to encourage exponential growth. When you go into GenBank, you see a gene and you can find information based solely on gene sequence, and maybe work your way into the protein sequence, and maybe even into an X-ray structure if something is available.

Austin says this initiative will lower the boundaries someone who is a molecular biologist would have to face if they want to understand how small molecules interact with their targets. "It's going to be a whole new way of educating people in areas that traditionally were difficult for them to understand."

"I have often wondered where the postgenomic next step is going to land," says DPI's Livingston. "I think those of us involved with this project believe that this has the opportunity to be an absolutely critical piece." 